

Non-Conservative Diffusion and its Application to Social Network Analysis

Rumi Ghosh, Kristina Lerman, Tawan Surachawala

Konstantin Voevodski

Shanghua Teng

*USC Information Sciences Institute
Marina del Rey, CA 90292, USA
{rumig,lerman,tawans}@isi.edu*

*Department of Computer Science University of Southern California
Boston University Los Angeles, CA 90007, USA
kvodski@gmail.com shanghua@usc.edu*

Abstract—Is the random walk appropriate for modeling and analyzing social processes? We argue that many interesting social phenomena, including epidemics and information diffusion, cannot be modeled as a random walk, but instead must be modeled as broadcast-based or non-conservative diffusion. To produce meaningful results, social network analysis algorithms have to take into account differences between these diffusion processes. We formulate conservative (random walk-based) and non-conservative (broadcast-based) diffusion mathematically and show how these are related to well-known metrics: PageRank and Alpha-Centrality respectively. This formulation allows us to unify two distinct areas of network analysis — centrality and epidemic models — and leads to insights into the relationship between diffusion and network structure, specifically, the existence of an epidemic threshold in non-conservative diffusion. We demonstrate, by ranking nodes in an online social network used for broadcasting news, that non-conservative Alpha-Centrality leads to a better agreement with empirical ranking schemes than conservative PageRank. In addition, we give a scalable approximate algorithm for computing the Alpha-Centrality in a massive graph. We hope that our investigation will inspire further exploration of the applications of non-conservative diffusion in social network analysis.

Keywords—social networks; centrality; diffusion

I. INTRODUCTION

Social network analysis algorithms examine the topology of the network to identify central nodes within it or groups of tightly connected nodes. In many cases, these algorithms make implicit assumptions about the underlying diffusion process taking place on the network [1]. Some of the best-known algorithms used for graph partitioning [2] and ranking, including PageRank and its variants [3], [4], are based on the random walk [5], [6]. A random walk on a graph is a stochastic process which starts at some node, and at each time step randomly selects one of the neighbors of the current node. The random walk is used to model chemical diffusion and other physical processes in which the total amount of the diffusing substance remains constant. However, the random walk may not be appropriate for modeling phenomena of greatest interest to social scientists, including adoption of innovation [7], [8], the spread of epidemics [9], [10] and word-of-mouth recommendations [11], viral marketing campaigns [12], [13], and information diffusion [14].

These examples are modeled as contact processes, where an activated or “infected” node activates its neighbors with some probability. Rather than picking one of the neighbors, in these stochastic processes each node *broadcasts* to all its neighbors. Therefore, unlike the random walk, which conserves the amount of substance diffusing on the network, contact processes are fundamentally non-conservative. When an idea, information, or disease spreads from one individual to her neighbors, the amount of information or disease changes (Chapter 5, [15]). If the random walk cannot model these social processes, can we trust results of social network analysis algorithms that are based on the random walk? If not, what are the appropriate metrics and methods to use for network analysis? And how can we empirically evaluate their performance?

In this paper we present a mathematical formulation of conservative and non-conservative diffusion and demonstrate how these are related to two well-known centrality metrics used to rank nodes in a network: PageRank [3] and Alpha-Centrality [16]. While PageRank is known to be equivalent to conservative diffusion [5], [6], we show that Alpha-Centrality is related to non-conservative diffusion, of which epidemic models are the best known example. Our formulation unifies two distinct research areas within network analysis — centrality measures and epidemic models — and leads to insights into relationship between dynamic processes and network structure. One consequence of the analysis is the existence of a threshold, called epidemic threshold [17], below which non-conservative diffusion dies out, but above which it reaches significant fraction of nodes within the network. We elucidate connection between the properties of Alpha-Centrality and the location of the epidemic threshold.

We demonstrate empirically that the choice of the centrality metric impacts our ability to identify central or influential nodes within a network. Specifically, we study online social network of Digg involved in spreading news stories. The spread of news on Digg can be modeled as an epidemic process [18], and hence represents non-conservative diffusion. One benefit of using social media data sets is that user activity on these sites provides an independent measure of influence. We define two empirical measures of influence that serve as the ground truth for ranking

users within this social network. We compare the rankings produced by different centrality metrics to the ground truth and show that non-conservative Alpha-Centrality leads to a better agreement with the ground truth than conservative PageRank. Finally, we present an approximate algorithm that can efficiently compute Alpha-Centrality for massive graphs and give a proof of its performance guarantees.

Specifically, the paper makes the following contributions:

- Define and classify diffusion processes occurring on networks (Section II).
- Establish a connection between diffusion and network structure (Section III). We also show how centrality metrics are related to diffusion processes occurring on the network.
- Empirically validate the hypothesis that non-conservative metric better predicts central people in an online social network used for (non-conservative) information diffusion than a conservative metric (Section IV).
- Provide a fast approximate algorithm to compute Alpha-Centrality (Section V).

II. CLASSES OF DIFFUSION PROCESSES

We represent a network by a directed, weighted graph $G = (V, E)$ with V nodes and E edges. We use $w[u, v]$ to specify the weight of the edge from u to v . The adjacency matrix of the graph is defined as: $A[u, v] = w[u, v]$ if $(u, v) \in E$; otherwise, $A[u, v] = 0$. $N(u)$ is the set of out-neighbors of u : $N(u) = \{v \in V | (u, v) \in E\}$, $d_{\text{out}}(u)$ is the out-degree of u : $d_{\text{out}}(u) = \sum_{v \in N(u)} w[u, v]$, and d_{max} is the maximum out-degree of any node in the graph. Note that the L_1 -norm of any argument is given by $\|\cdot\|_1$.

Network diffusion is a dynamic stochastic process that distributes some quantity, which we generically refer to as weight, on a network or a graph. Diffusion process is described mathematically by a function $F : (R^+ \cup \{0\})^{|V|} \rightarrow (R^+ \cup \{0\})^{|V|}$, i.e., a map from a $|V|$ -dimensional non-negative vector to a $|V|$ -dimensional non-negative vector (here V is the number of nodes). The vector $x \in (R^+ \cup \{0\})^{|V|}$ represents the weight each node has at time t . The function $F(x)$ maps the weight vector at time t to the weight vector at time $t + 1$.

A. Conservative Diffusion

We call a stochastic process $\mathcal{C}_t : (R^+ \cup \{0\})^{|V|} \rightarrow (R^+ \cup \{0\})^{|V|}$ that simply redistributes the weights among the nodes of the graph, with the total weight remaining constant, *conservative diffusion*. In other words, in conservative diffusion for all $x \in (R^+ \cup \{0\})^{|V|}$, $\|x\|_1 = \|\mathcal{C}_t(x)\|_1$.

To motivate our mathematical formulation of conservative diffusion, we imagine a hypothetical society where each member has some amount of money to redistribute. If money cannot be created or destroyed, money redistribution represents a conservative diffusion process. Let $\mathcal{X}_c(t)$ be

the vector representing the amount of money each member has at time t , and $\Delta(t)$ represent the amount they receive at time t . We consider a distribution process where the amount *redistributed* at each step, depends on the money each member *received* in the previous step. We focus on this redistribution process, because, as we show later, this is the process underlying popular network models. A different conservative process could be one in which the amount redistributed in each step depends on the amount each member *had* in the previous time step. This would lead to a different mathematical formulation of the diffusion process.

At time $t + 1$, each member retains a fraction $(1 - \alpha)$, with $0 \leq \alpha \leq 1$, of this amount and distributes the rest among its neighbors. Let \mathcal{W}_c be the *transfer matrix*, with $\mathcal{W}_c[p, q]$ representing the fraction of the amount to be redistributed by node p transferred to q . Therefore, the amount of money nodes receive at time $t + 1$ via redistribution can be written as:

$$\Delta(t + 1) = \alpha \Delta(t) \mathcal{W}_c.$$

Thus the transfer matrix encodes the rules of diffusion. If each member *divides* $\alpha \Delta(t)$ equally amongst her out-neighbors, then $\mathcal{W}_c = D^{-1}A$, where the degree matrix D is a diagonal matrix of out-degrees, and A is the adjacency matrix.

Step by step, conservative diffusion looks as follows. Initially, at time $t = 0$, let the weight each node receives be $\Delta(0) = \mathcal{X}_c(0)$. Let the process begin at time $t = 1$, when each node keeps $(1 - \alpha)$ of that amount and divides the rest ($\alpha \Delta(0)$) evenly between its out-neighbors. The amount that out-neighbors receive from redistribution at time $t = 1$ is $\Delta(1) = \alpha \Delta(0) \mathcal{W}_c = \alpha \mathcal{X}_c(0) \mathcal{W}_c$.

At time $t = 2$, each node retains $(1 - \alpha)$ of the amount $\Delta(1)$ it received at time $t = 1$, and divides the rest among its out-neighbors. Therefore, the amount received by the out-neighbors is $\Delta(2) = \alpha \Delta(1) \mathcal{W}_c = \alpha^2 \mathcal{X}_c(0) \mathcal{W}_c^2$.

Continuing with this process further, at any time $t > 0$, each nodes retains $(1 - \alpha)$ of the amount of it received at time $t - 1$,

$$\begin{aligned} (1 - \alpha) \Delta(t - 1) &= (1 - \alpha) \alpha \Delta(t - 2) \mathcal{W}_c \\ &= (1 - \alpha) \alpha^{t-1} \mathcal{X}_c(0) \mathcal{W}_c^{t-1}, \end{aligned} \quad (1)$$

and divides the rest among her out-neighbors. Hence, the amount received by the out-neighbors is

$$\Delta(t) = \alpha \Delta(t - 1) \mathcal{W}_c = \alpha^t \mathcal{X}_c(0) \mathcal{W}_c^t. \quad (2)$$

The total weight (or amount of money in our example) the nodes have at time t , $\mathcal{X}_c(t)$, is the amount they retained from all previous time steps and the amount they receive

from in-neighbors at time t :

$$\begin{aligned}
\mathcal{X}_c(t) &= (1 - \alpha) \sum_{k=0}^{t-1} \Delta(k) + \Delta(t) \\
&= \sum_{k=0}^{t-1} (1 - \alpha) \alpha^k \mathcal{X}_c(0) \mathcal{W}_c^k + \alpha^t \mathcal{X}_c(0) \mathcal{W}_c^t \\
&= (1 - \alpha) \mathcal{X}_c(0) + \alpha \mathcal{X}_c(t-1) \mathcal{W}_c.
\end{aligned} \tag{3}$$

As $t \rightarrow \infty$, this equation reduces to

$$\begin{aligned}
\mathcal{X}_c(t \rightarrow \infty) &= (1 - \alpha) \mathcal{X}_c(0) + \alpha \mathcal{X}_c(t \rightarrow \infty) \mathcal{W}_c \\
&= (1 - \alpha) \mathcal{X}_c(0) (I - \alpha \mathcal{W}_c)^{-1}
\end{aligned} \tag{4}$$

The transfer matrix \mathcal{W}_c is a stochastic matrix, since its rows sum up to 1. If, as described above, the weight to be redistributed at each step is divided equally between the out-neighbors, then $\mathcal{W}_c = D^{-1}A$. However, if instead each node decides to keep a portion δ of this amount, this leads to a more general form of the transfer matrix:

$$\mathcal{W}_c = \delta I + (1 - \delta) D^{-1}A. \tag{5}$$

Note that in our hypothetical society, the total amount of money remains constant: if $\mathcal{C}_t : \mathcal{X}_c(0) \rightarrow \mathcal{X}_c(t)$ defines a diffusion process, then $\|\mathcal{X}_c(0)\|_1 = \|\mathcal{C}_t(\mathcal{X}_c(0))\|_1$. Hence this is a *conservative diffusion process*. In the above scenario, \mathcal{C}_t is a linear mapping; therefore, we call the diffusion processes given by Eqs. 3 and 4 *linear conservative diffusion*. In a more general representation, \mathcal{C}_t can even be a non-linear mapping, describing non-linear conservative diffusion.

Random Walk as Conservative Diffusion: Like money transfer, a random walk on a graph can be modeled as a conservative diffusion process, since the probability to find a random walker on any node of the graph is always one. A random walk with random jumps or restarts can be described mathematically as follows. Let the initial probability to find the random walker on any node be uniform, i.e., $\mathcal{X}_c(0)[i] = \frac{1}{|V|}$. At any time t , with probability α the random walker at node i chooses one of the neighbors of i uniformly at random and jumps to it. With probability $(1 - \alpha)$, it chooses any node on the graph uniformly at random and jumps to it. Let matrix X encode the probability of jumping to any node, $X[i, j] = \frac{1}{|V|}$, and $\mathcal{W}_c = D^{-1}A$. Then the probability of finding the random walker at node j at time t is given by

$$\begin{aligned}
\mathcal{X}_c(t) &= (1 - \alpha) \mathcal{X}_c(t-1) X + \alpha \mathcal{X}_c(t-1) \mathcal{W}_c \\
&= (1 - \alpha) \mathcal{X}_c(0) + \alpha \mathcal{X}_c(t-1) \mathcal{W}_c.
\end{aligned}$$

This is exactly the same as Eq. 3. Therefore, a random walk with a uniform starting vector is mathematically equivalent to a linear conservative diffusion process.

B. Non-Conservative Diffusion

A diffusion process where the total weight can change in time is a *non-conservative diffusion process*. Formally, a function $\mathcal{N}_t : (R^+ \cup \{0\})^{|V|} \rightarrow (R^+ \cup \{0\})^{|V|}$ defines a *non-conservative diffusion process* if for some $x \in (R^+ \cup \{0\})^{|V|}$, $\|x\|_1 \neq \|\mathcal{N}_t(x)\|_1$.

To illustrate the difference between conservative and non-conservative processes, we return to our hypothetical society. Again, imagine that each member has some amount of money, however, unlike the previous example, each member also has a money printing machine, so that instead of dividing the money she receives equally between her out-neighbors, she can give each neighbor the same amount by printing extra money as needed.

Let $\Delta(t)$ be the vector representing the amount of money each member receives at time t . At the next time step, each member prints a fraction α of this amount to give to each of her out-neighbors. The additional amount that she produces for her out-neighbors can be expressed using the *replication matrix* $\mathcal{W}_n = A$. Therefore, $\Delta(t+1) = \alpha \Delta(t) \mathcal{W}_n$.

Initially, let $\Delta(0) = \mathcal{X}_n(0)$. At time $t = 1$, each member prints $\alpha \Delta(0)$ for each of her out-neighbors:

$$\Delta(1) = \alpha \Delta(0) \mathcal{W}_n = \alpha \mathcal{X}_n(0) \mathcal{W}_n.$$

Similarly, at time $t = 2$,

$$\Delta(2) = \alpha \Delta(1) \mathcal{W}_n = \alpha^2 \mathcal{X}_n(0) \mathcal{W}_n^2.$$

Continuing this process, additional amount of money each member produces or receives at time t is:

$$\Delta(t) = \alpha \Delta(t-1) \mathcal{W}_n = \alpha^t \mathcal{X}_n(0) \mathcal{W}_n^t \tag{6}$$

Therefore, the total amount that each member has at time t is obtained by summing up the additional amount she accrues or receives from her in-neighbors at each time step:

$$\begin{aligned}
\mathcal{X}_n(t) &= \sum_{k=0}^t \Delta(k) = \sum_{k=0}^t \mathcal{X}_n(0) (\alpha \mathcal{W}_n)^k \\
&= \mathcal{X}_n(0) + \alpha \mathcal{X}_n(t-1) \mathcal{W}_n
\end{aligned} \tag{7}$$

At time $t \rightarrow \infty$, Eq. 7 reduces to

$$\mathcal{X}_n(t \rightarrow \infty) = \mathcal{X}_n(0) \sum_{k=0}^{t \rightarrow \infty} (\alpha \mathcal{W}_n)^k \tag{8}$$

which can be solved to yield

$$\begin{aligned}
\mathcal{X}_n(t \rightarrow \infty) &= \mathcal{X}_n(0) + \mathcal{X}_n(t \rightarrow \infty) (\alpha \mathcal{W}_n) \\
&= \mathcal{X}_n(0) (I - \alpha \mathcal{W}_n)^{-1}.
\end{aligned} \tag{9}$$

This expression is defined for $\alpha < 1/\lambda_1$, where λ_1 is the largest eigenvalue, or spectral radius, of \mathcal{W}_n .

More generally, if along with producing α of what it receives from each of its in-neighbors, a node also produces

a portion δ of this amount for itself, this results in a more general form of the replication matrix:

$$\mathcal{W}_n = \frac{\delta}{\alpha}I + A. \quad (10)$$

The diffusion process defined by Eqns. 7–9 is non-conservative, since $\|\mathcal{X}_n(0)\|_1 \neq \|\mathcal{N}_t(\mathcal{X}_n(0))\|_1$. Moreover, it is linear, although the function \mathcal{N}_t may also be non-linear.

We can model non-conservative diffusion as a random walk with birth, where at each time step, the random walker gives birth to one or more new walkers. The number of random walkers on the network, therefore, will change with time. Several social phenomena can be modeled using this framework. In rumor propagation, for example, some information spreads in a community as people pass it to their neighbors. This process is non-conservative, since the number of informed individuals grows in time. We can model rumor propagation as a random walk on the friendship graph, where the random walker (rumor) randomly selects one of the neighbors of the informed node to move to, while leaving a clone of itself at the node. Cloning is required for the node to remain informed. If the informed node immediately forgot the rumor (no cloning required), then rumor propagation could be modeled by a simple random walk and would be conservative in nature, since the number of informed individuals would always be one.

Epidemics as Non-Conservative Diffusion: Non-conservative diffusion provides a useful framework for thinking about epidemics and other spreading processes and leads to insights into the relation between network structure and dynamics of spreading processes. In a spreading process, information or virus spreads from an informed or infected individual to her network neighbors. In order to model a spreading process accurately, the structure of the underlying network has to be taken into account. Wang *et al.* [17] modified existing SIS models [19] to take network structure into account in order to describe the spread of epidemics in real networks. We demonstrate that this model is equivalent to the *linear non-conservative diffusion process* (Equation 7).

Consider a virus spreading on a network, where at each time step, a node infected with the virus may infect its out-neighbors with probability μ (virus birth rate). At each time step, an infected node may also be cured with probability β (virus curing rate). Wang *et al.* [17] showed that the probability $p_{i,t}$ that node i is infected at time t can be written in matrix notation as

$$P_t = P_{t-1}((1 - \beta)I + \mu A) = P_0((1 - \beta)I + \mu A)^t$$

where P_t is a vector $(p_{1,t}, p_{2,t}, \dots)$, and P_0 is the initial probability of infection.¹ This formulation makes

¹This model holds true only when $p_{i,t}$ is very small and there may be situations where $p_{i,t} > 1$. Therefore a more accurate interpretation is that the probability of infection is proportional to $p_{i,t}$.

the probability of infection at time t , P_t , exactly equal to the additional weight, $\Delta(t)$, accrued at each step in non-conservative diffusion, as shown in Eq. 6 with the replication matrix $\mathcal{W}_n = \frac{1-\beta}{\mu}I + A$ and $\alpha = \mu$. In the model described above, there exists an epidemic threshold τ such that for $\mu/\beta < \tau$ epidemic will die out, and $\mu/\beta > \tau$ it will spread to a significant fraction of nodes [17]. For any graph, this threshold is given by the inverse of the largest eigenvalue of the graph's adjacency matrix A : $\tau = 1/|\lambda_1|$.

III. DIFFUSION AND NETWORK STRUCTURE

The complex interplay between network structure and diffusion has broad implications for modeling and understanding networks. While it is known that the macroscopic properties of diffusion (e.g., epidemic threshold) are affected by network structure [17], [20], the impact of diffusion on our understanding of network structure is less appreciated. In this paper we show that social network analysis, specifically, identifying central or influential nodes, is affected by the characteristics of the diffusion process occurring on the network. Centrality metrics used for this task examine the topology of the network only. However, these metrics usually make implicit assumptions about the nature of diffusion process taking place on the network [1], with each metric leading to a different, even conflicting notion, of who the central nodes are. We show that the characteristics of network diffusion should be one of the guiding principles in choosing an appropriate network analysis algorithm.

A. Centrality and Diffusion

A node's centrality predicts its relative importance, influence, or prestige within the network. Over the years many different centrality metrics have been introduced for social network analysis, including degree centrality, betweenness centrality [21], eigenvector centrality [22], PageRank [3] and Alpha-Centrality [23].

1) *Page Rank:* A PageRank vector $\text{pr}_\alpha(s, t)$ is the steady state probability distribution of a random walk with damping factor α (restart probability = $1 - \alpha$). The starting vector s , gives the probability distribution for where the walk transitions after restarting. The transfer matrix encodes the transition probabilities of a random walk on the network, $W = D^{-1}A$. PageRank is the unique steady state solution $\text{pr}_\alpha(s, \infty)$ of:

$$\text{pr}_\alpha(s, t) = (1 - \alpha)s + \alpha \text{pr}_\alpha(s, t - 1)W \quad (11)$$

For ease of convention, we denote PageRank by $\text{pr}_\alpha(s)$. Hence

$$\text{pr}_\alpha(s) = (1 - \alpha)s + \alpha \text{pr}_\alpha(s)W \quad (12)$$

Equation 12 is identical to the steady state solution of the linear conservative diffusion process given by Eq. 4 where $W = W_c = D^{-1}A$ and $s = \mathcal{X}_c(0)$. Therefore, *PageRank is the steady state solution of conservative diffusion*, and PageRank is a conservative metric. Most of the other metrics

derived from the random walk make an implicit assumption of conservative diffusion taking place on a network.

2) *Alpha-Centrality*: Alpha-Centrality [23] measures the total number of paths from a node, exponentially attenuated by their length. For a starting vector s and attenuation parameter α , the Alpha-Centrality vector is the steady state solution to:

$$\text{cr}_\alpha(s, t) = s + \alpha \text{cr}_\alpha(s, t-1)A. \quad (13)$$

The starting vector s is usually taken as in-degree centrality [23]. For ease of convention, we shall denote $\text{cr}_\alpha(s, t \rightarrow \infty)$ by $\text{cr}_\alpha(s)$. As $t \rightarrow \infty$, the solution converges to

$$\text{cr}_\alpha(s) = s + \alpha \text{cr}_\alpha(s)A, \quad (14)$$

which holds while $|\alpha| < \frac{1}{|\lambda_1|}$.

One difficulty in applying Alpha-Centrality in network analysis is that its key parameter α is bounded by λ_1 , the spectral radius of the network. As a result, the metric diverges at this value of the parameter. To overcome this, *normalized Alpha-Centrality* [24] has been recently introduced, which we denote by $\text{ncr}_\alpha(s, t)$. It normalizes the score of each node by the sum of the Alpha-Centrality scores of all the nodes. The new metric avoids the problem of bounded parameters while retaining the desirable characteristics of Alpha-Centrality, namely its ability to differentiate between local and global structures.

Normalized Alpha-Centrality $\text{ncr}_\alpha(s, t \rightarrow \infty)$ is defined using the system of equations shown below:

$$\text{ncr}_\alpha(s, t) = \frac{1}{\|\text{cr}_\alpha(s, t)\|_1} \text{cr}_\alpha(s, t) \quad (15)$$

The new metric is well defined for $\alpha \geq 0$ ($\alpha \neq \frac{1}{|\lambda_1|}$).

Equation 13 and Eq. 15 are mathematically equivalent to Eq. 8, with starting vector $\mathcal{X}_n(0) = c \cdot s$, where $c = 1$ for Alpha-Centrality and

$$c = \frac{1}{\sum_{i,j} \sum_{k=0}^t \alpha^k A^k[i, j]}$$

for normalized Alpha-Centrality. Therefore, *Alpha-Centrality is a steady state solution of linear non-conservative diffusion* and is a non-conservative metric. Other non-conservative metrics include degree centrality, Katz score [25], SenderRank [26], and eigenvector centrality [16].

B. Length Scales and Epidemic Threshold

The link between Alpha-Centrality (and normalized Alpha-Centrality) and non-conservative diffusion leads to a fundamental insight into the relationship between network structure and the size of epidemics. Let us look more carefully at Equation 8. The weight distribution given by this equation depends on the initial weight distribution ($\mathcal{X}_n(0)$) and the power series of matrices $S(\alpha, t) = \sum_{k=0}^t (\alpha \mathcal{W}_n)^k$. For illustrative purposes, we can interpret \mathcal{W}_n to be the

adjacency matrix of some graph G' . Then each element in the power series $S(\alpha, t)[i, j]$ can be interpreted as the number of attenuated paths from node i to node j up to length t in that graph G' . In Alpha-Centrality or normalized Alpha-Centrality, these paths determine the centrality of the node along with the initial distribution of weights. The probability of non-conservative diffusion reaching node j from i through a path of length k is α^k . $S(\alpha, t)[i, j]$ then characterizes the expected number times a non-conservative diffusion process initiated at node i reaches node j up until time t . For example, let node i be infected by a virus and initiate a viral infection in the network. If viral infection can be modeled as linear non-conservative diffusion (Section II-B), the probability that node j will get infected by the viral infection from node i through a path of length k would be α^k . Then $S(\alpha, t)[i, j]$ would quantify the expected number of viruses reaching node j when the viral infection is initiated at node i .

As shown in the Appendix, the expected path length of diffusion as $t \rightarrow \infty$, is $\frac{1}{1-\alpha\lambda_1}$ if $\alpha < \frac{1}{|\lambda_1|}$ and $O(t)$ if $\alpha > \frac{1}{|\lambda_1|}$. Therefore, $\frac{1}{|\lambda_1|}$ is a *threshold*: for α below threshold, the expected path length converges with time, while for α above the threshold, it diverges. Note that this threshold is equivalent to the epidemic threshold (Section II-B). Thus from the diffusion point of view, given the network structure and nature of diffusion, α (for $\alpha < 1/|\lambda_1|$) determines how far, on average, a node's effect will be felt and sets the length scale of the interaction. When α is small, Alpha-Centrality or normalized Alpha-Centrality probes only the local structure of the network. As α grows, structurally longer paths become more important, (normalized) Alpha-Centrality becomes a global measure and the weight diffuses to a greater number of nodes.

C. Choosing the Centrality Metric

When applied to the same network, different centrality metrics may lead to different, often incompatible, views of who the important nodes are. We illustrate these differences on a toy network shown in Fig. 1, where a link from node u to node v indicates that node v is an out-neighbor of u , e.g., u is a follower of v in an online social network.

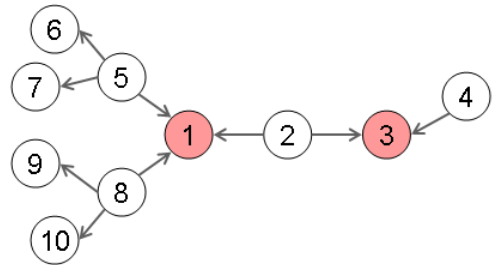


Figure 1. An example network, where node 1 has the highest Alpha-Centrality followed by node 3. In contrast node 3 has the highest PageRank followed by node 1.

Even in this simple example, PageRank and Alpha-Centrality disagree about who the most important node is. PageRank without restarts ranks node 3 highest, followed by node 1. In contrast, Alpha-Centrality ranks node 1 above node 3. The difference in rankings produced by the two centrality metrics is due to the difference in the underlying diffusion process that redistributes the weights of the nodes. Assume that all nodes start with equal weights, which then evolve according to the rules of diffusion. In PageRank without restarts (damping factor $\alpha = 1$), each follower divides its weight equally among its d_{out} out-neighbors, and hence transfers a fraction $1/d_{out}$ to each. Thus, node 5 contributes $1/3$ of its weight to node 1, and so will node 8. Node 3, on the other hand, will get the entire weight of node 4, giving it a higher weight than node 1 and therefore, a higher rank.

In contrast to PageRank, Alpha-Centrality has nodes update their weights by copying a portion of their followers' weights. For consistency with PageRank, we take $\alpha = 1$. Thus, node 1 will receive the entire weights of nodes 2, 5 and 8, while node 3 will only receive the weights from nodes 2 and 4. Therefore, the weight of node 1 will be greater than node 3, and consequently, it will be ranked higher by Alpha-Centrality.

Which ranking is right? How do we choose the right centrality metric for our problem? We claim that the choice of the centrality metric has to be motivated by details of the diffusion process taking place on the network. To analyze networks on which processes such as random walk, web surfing, money and used goods exchange are taking place, conservative metrics, such as PageRank, are appropriate. On the other hand, to study social networks on which information or epidemics are spreading, non-conservative metrics, such as Alpha-Centrality, should be used. In other words, *the centrality metric that best predicts important nodes in a network is one whose implicit dynamics most closely matches the diffusion process occurring on the network.*

IV. PREDICTING INFLUENTIALS IN ONLINE SOCIAL NETWORKS

Online social networks on sites such as Facebook, Twitter, and Digg have become important hubs of social activity and conduits of information. The ever-growing popularity of these networks and overwhelming amount of information contained in them, necessitates the need for a more principled approach to social network analysis and data mining. Correctly identifying influential nodes on these networks can have far-reaching consequences for identifying noteworthy content [27], targeted information dissemination [12], and other applications. While a variety of methods [28], [29] have been used to identify influential users in online social networks, each metric leads to a different result, and no justification for these metrics have been proposed.

Fortunately, by exposing activity of their users, online social networks provide a unique opportunity to study dy-

namic processes on networks. We analyze information flow on the social news aggregator Digg and use this data to empirically evaluate centrality metrics. By posting a story on Digg, submitter broadcasts it to her followers. When another user votes for this story, she broadcasts it to her own followers. We claim that since broadcast-driven information diffusion on Digg is non-conservative in nature, a non-conservative metric will better identify influential users than a conservative metric.

The Digg dataset comprises around 300K users and over 1 million friendship links, from which we can extract the directed follower network of active users. These users were active in spreading stories on Digg by either submitting them or voting for them, since both activities expose the story to the submitter or voter's followers. The data set contains more than 3 million votes on more than 3000 stories promoted to Digg's front page in June 2009. Note that the underlying follower graph was extracted separately of user activity. In fact, user activity provides an independent measure of influence in online social networks that we use to evaluate the centrality metrics.

A. Empirical Estimates of Influence

Katz and Lazarsfeld [30] defined influentials as "individuals who were likely to influence other persons in their immediate environment." In the years that followed, many attempts were made to identify people who influenced others to adopt a new practice or product by looking at how innovations or word-of-mouth recommendations spread [31]. The rise of online social networks has allowed researchers to trace the flow of information through social links on a massive scale. Using the new empirical foundation, some researchers proposed to measure a person's influence by the size of the cascade he or she triggers [12]. However, as Watts and Dodds [32] note, "the ability of any individual to trigger a cascade depends much more on the global structure of the influence network than on his or her personal degree of influence." Alternatively, Trusov *et al.* [33] defined influential people in an online social network as those whose activity stimulates those connected to them to increase their activity, while Cha *et al.* [28] used the number of retweets and mentions to measure user influence on Twitter.

Motivated by these works, we measure influence by analyzing users' activity on an online social network. Suppose some user, the *submitter*, posts a new story on Digg. We measure the activity submitter's post generates by the number of times it is re-broadcast by followers. Whether or not a user will re-broadcast the story depends on (i) story quality and (ii) influence of the submitter. We assume that story's quality is uncorrelated with the submitter.² Therefore, we can average out its contribution to the activity a submitter generates by aggregating over all stories submitted by the

²This is a fairly strong assumption, but it appears to hold at least for Digg [27].

same user. We claim that the residual difference between submitters can be attributed to variations in influence. We propose two metrics to measure submitter’s influence: (i) average number of follower votes her posts generate and (ii) average size of the cascades her posts trigger.

B. Comparison of Centrality Metrics

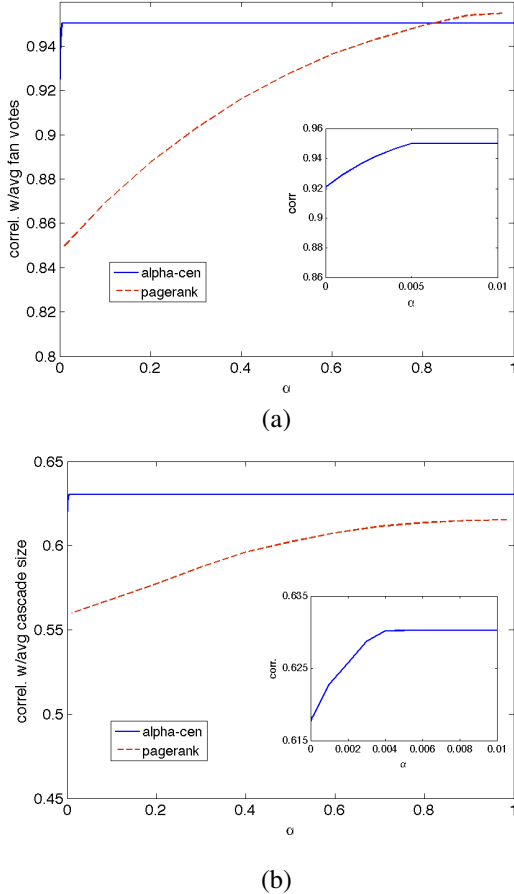


Figure 2. Correlation between the rankings produced by the empirical measures of influence and those predicted by normalized Alpha-Centrality and PageRank. We use (a) the average number of follower votes and (b) average cascade size as the empirical measures of influence. The inset zooms into the variation in correlation for $0 \leq \alpha \leq 0.01$

We use the empirical estimates of influence to rank a subset of users in our sample who submitted more than one story which received at least 100 votes. There were 289 Digg users in this sample. We used the rankings produced by either empirical estimate as the ground truth to evaluate the performance of different centrality metrics. We studied PageRank (with uniform starting vector) and normalized Alpha-Centrality, both of which were computed considering the entire Digg follower network as a graph, with the thousands of users as nodes and the millions of friendship links as edges. We use Pearson’s correlation coefficient (since ties in rank may exist) to compare the rankings predicted

by the different centrality metrics with the ground truth. Figure 2 shows how the correlation in rankings changes with the parameter $0 \leq \alpha \leq 1$. This parameter stands for the *attenuation factor* for normalized Alpha-Centrality (see Equation 13) and the *damping factor* (*restart probability* $= 1 - \alpha$) for PageRank (see Equation 11). If we used Alpha-Centrality instead of normalized Alpha-Centrality, we would have been bounded by its formalization, to compute the rankings only for $\alpha < \frac{1}{|\lambda_1|}$. Note that the correlation of PageRank at $\alpha = 0$ (restart probability $= 1$) with the empirical estimate cannot be computed because standard deviation of PageRank rankings would be zero in this case. Various studies have tested different damping factors for Page Rank, but it is generally assumed that the damping factor should be set around $\alpha = 0.85$ [3]. Boldi et al. [34] claim that in case of PageRank, “for real-world graphs values of α close to 1 do not give a more meaningful ranking.” Except for values α close to 1, the influence rankings calculated from Alpha-Centrality correlated better with the empirical estimates of influence rankings than PageRank rankings. Therefore, we conclude that Alpha-Centrality predicts central users in the Digg social network better than PageRank.

V. APPROXIMATION ALGORITHM FOR ALPHA-CENTRALITY

In order to compute the exact Alpha-Centrality vector we have to solve Equation 13, which requires us to compute a matrix inverse. Computing a matrix inverse in a naive implementation, takes $O(n^3)$ time (where n is the number of nodes in the network), so this is difficult to compute for large networks. One way to compute an approximate solution is to use the alternate formulation given in Equation 7, and compute $s(I + \alpha A + \alpha^2 A^2 + \alpha^3 A^3 + \dots)$, until the α^i coefficient grows sufficiently small. While this technique is effective in practice, computing A^i in each iteration, using a naive implementation would have must take at least n^2 time, and it is not clear how many iterations we need to get a good approximation. In this section we present an algorithm for approximating Alpha-Centrality, which has a single parameter that controls both the runtime and the quality of the produced approximation.

A description of our algorithm is given in Algorithm 1. Our procedure is similar to the algorithm for approximating PageRank that is given in [2]. Our algorithm takes the network, the starting vector s , α , and an approximation parameter δ ($0 < \delta \leq 1$) as input, and computes an approximate Alpha Centrality vector where each entry has error of at most δ (see Theorem 2). In order to approximate a centrality vector with starting vector s , we maintain an *approximate centrality* vector \tilde{c} and a *residual* vector r . Initially r is equivalent to the starting vector s ; the algorithm iteratively moves content from r to \tilde{c} until each entry in r is small.

When the α parameter is fixed, we use $\text{cr}(s)$ to denote $\text{cr}_\alpha(s)$. We will also use $[\text{cr}(s)](u)$ to refer to how much

Algorithm 1 Approximate-Centrality(V, E, s, α, δ)

```

1:  $\epsilon = \delta \|s\|_1 / n$ ;
2:  $r = s$ ;
3: Queue  $q = \text{new Queue}()$ ;
4: for each  $u \in V$  do
5:    $\tilde{c}r(u) = 0$ ;
6:   if  $r(u) > \epsilon$  then
7:      $q.\text{add}(u)$ ;
8:   end if
9: end for
10: while  $q.\text{size}() > 0$  do
11:    $u = q.\text{dequeue}()$ ;
12:    $\tilde{c}r(u) = \tilde{c}r(u) + r(u)$ ;
13:    $T = \alpha \cdot r(u)$ ;
14:    $r(u) = 0$ ;
15:   for each  $v \in N(u)$  do
16:      $r(v) = r(v) + T \cdot w(u, v)$ ;
17:     if  $!q.\text{contains}(v)$  and  $r(v) > \epsilon$  then
18:        $q.\text{add}(v)$ ;
19:     end if
20:   end for
21: end while
22: return  $\tilde{c}r$ ;

```

content vertex u has in $\text{cr}(s)$. We give our formal performance guarantee for Algorithm 1 in Theorem 2. This performance guarantee is based on Lemma 1, which shows that in any step of the algorithm, the approximate centrality computed for Alpha-Centrality with s as starting vector, is always exactly equivalent to Alpha Centrality with $s - r$ as starting vector, where r is the residual vector in that step i.e. throughout the execution of the algorithm, the error in the approximate centrality vector is dependent on the amount of content remaining in the residual vector.

Our arguments depend on the linearity of the centrality computation with respect to the starting vector, which is easy to verify. We can show that $\text{cr}_\alpha(s_1) + \text{cr}_\alpha(s_2) = \text{cr}_\alpha(s_1 + s_2)$, and $c \cdot \text{cr}_\alpha(s) = \text{cr}_\alpha(c \cdot s)$.

Lemma 1: The invariant $\tilde{c}r = \text{cr}(s - r)$ is maintained throughout the execution of the while-loop.

Proof: Before the loop starts, we have $r = s$ and $\tilde{c}r = \vec{0}$, so $\text{cr}(s - r) = \text{cr}(\vec{0}) = \vec{0} = \tilde{c}r$. We can also show that if $\tilde{c}r = \text{cr}(s - r)$ holds prior to an iteration of the loop, then $\tilde{c}r' = \text{cr}(s - r')$ is still true after the iteration, where $\tilde{c}r'$ and r' are the updated approximate centrality and residual vectors.

We first observe that $\text{cr}(s)A = \text{cr}(sA)$. To see this, consider that by definition $\text{cr}(s) = s + \alpha \cdot \text{cr}(s)A$. Multiplying this equation by A we get $\text{cr}(s)A = sA + \alpha \cdot (\text{cr}(s)A)A$. This shows that $\text{cr}(s)A$ is by definition a centrality vector for starting vector sA . Moreover, we know that the solution to $\text{cr}(sA)$ is unique, so we have $\text{cr}(s)A = \text{cr}(sA)$. This observation shows that we can iteratively compute the centrality vector by expressing $\text{cr}(s)A$ as $\text{cr}(sA)$.

We will write the operations performed inside the while-loop using vector-matrix notation. We use e_u to denote a row vector that has all of its content in vertex u : $e_u(i) = 1$ if $i = u$; otherwise, $e_u(i) = 0$.

After an iteration of the loop we have $\tilde{c}r' = \tilde{c}r + r(u)e_u$, and $r' = r - r(u)e_u + \alpha r(u)e_u A$, where u is the vertex that is dequeued in line 11. We next specify the relationship between the approximate centrality and residual vectors before and after an iteration of the while-loop. Consider that

$$\begin{aligned}
\text{cr}(r) &= \text{cr}(r - r(u)e_u) + \text{cr}(r(u)e_u) \\
&= \text{cr}(r - r(u)e_u) + r(u)e_u + \text{cr}(\alpha r(u)e_u A) \\
&= \text{cr}(r - r(u)e_u + \alpha r(u)e_u A) + r(u)e_u \\
&= \text{cr}(r') + \tilde{c}r' - \tilde{c}r.
\end{aligned}$$

If $\tilde{c}r = \text{cr}(s - r)$, we have $\text{cr}(r) = \text{cr}(r') + \tilde{c}r' - \text{cr}(s - r)$. It follows that $\tilde{c}r' = \text{cr}(r) - \text{cr}(r') + \text{cr}(s - r) = \text{cr}(r - r' + (s - r)) = \text{cr}(s - r')$. ■

Theorem 2: Given an $\alpha \leq \frac{c}{d_{\max}}$ for some $c < 1$ and a uniform starting vector s , the vector $\tilde{c}r$ output by *Approximate-Centrality* satisfies $[\text{cr}(s)](u) \geq \tilde{c}r(u) \geq [\text{cr}(s)](u)(1 - \delta)$ for each vertex $u \in V$. The runtime of the algorithm is $O(\frac{n}{\delta} d_{\max})$.

Proof: Lemma 1 argues that $\tilde{c}r = \text{cr}(s - r) = \text{cr}(s) - \text{cr}(r)$ throughout the execution of the algorithm, so we have $\tilde{c}r(u) = [\text{cr}(s)](u) - [\text{cr}(r)](u)$ for all vertices $u \in V$. Given a uniform starting vector s , $s(u) = \|s\|_1 / n$ for all $u \in V$. The algorithm terminates when $r(u) \leq \epsilon$ for all $u \in V$, so we choose $\epsilon = \delta \cdot \|s\|_1 / n = \delta s(u)$ such that upon completion $r(u) \leq \delta s(u)$ for all $u \in V$.

Clearly, $[\text{cr}(s)](u) \geq \tilde{c}r(u)$ because r and $\text{cr}(r)$ are non-negative. We can also show that given that $r(u) \leq \delta s(u)$ for all $u \in V$, $[\text{cr}(r)](u) \leq \delta [\text{cr}(s)](u)$ for all vertices $u \in V$. It follows that $\tilde{c}r(u) = [\text{cr}(s)](u) - [\text{cr}(r)](u) \geq [\text{cr}(s)](u)(1 - \delta)$. Therefore we can see that indeed $[\text{cr}(s)](u) \geq \tilde{c}r(u) \geq [\text{cr}(s)](u)(1 - \delta)$ for all vertices $u \in V$.

We assume that α is chosen such that $\alpha \leq \frac{c}{d_{\max}}$ for some constant $c < 1$, where d_{\max} is the largest out-degree of any node in the graph. In order to bound the runtime of the algorithm, consider that each iteration of the while-loop decreases the sum of the entries of r by $(1 - \alpha \cdot d_{\text{out}}(u))r(u) > (1 - \alpha \cdot d_{\text{out}}(u))\epsilon \geq (1 - \alpha \cdot d_{\max})\epsilon \geq (1 - c)\epsilon$. Because $r = s$ at initialization and each iteration decreases $\|r\|_1$ by at least $(1 - c)\epsilon$, the number of iterations i must satisfy $i(1 - c)\epsilon \leq \|s\|_1$. Therefore the number of iterations may be at most $\frac{\|s\|_1}{(1 - c)\epsilon} = O(\|s\|_1 / \epsilon)$. The cost of each iteration is proportional to the out-degree of the node that is dequeued, so the worst-case runtime of the algorithm is $O(\|s\|_1 / \epsilon \cdot d_{\max})$. For our choice of ϵ this is equivalent to $O(\frac{n}{\delta} d_{\max})$. ■

A. Quality of Approximate Results

We compare the performance of the approximate algorithm with the power iteration method in Equation 13 using the indegree as the starting vector, like in [16] and

[22]. To compute Alpha-centrality using the approximate algorithm, we fix ϵ (Algorithm 1) to be 3.57×10^{-8} and 1.42×10^{-8} guaranteeing that the error in approximation would be less than 1% ($\delta < 0.01$). We terminate the power iteration algorithm after 100 iterations in Digg and 10 to 100 iterations in Twitter. We calculate the RMS (root mean square) error of the approximate algorithm with respect to the power iteration algorithm, for different values of α . The RMS error averaged over all values of α , is 0.797% and 0.75% for Digg and Twitter respectively.

VI. RELATED WORK

The interplay of the structural properties of the underlying network with the diffusion processes occurring in it, contributes to the complexity of real-life networks. For example in epidemiology, the dynamics of disease spread on a network and the epidemic threshold is closely related to its spectral radius of the graph [17]. Similarly, random walk on a graph is closely related Laplacian of the graph [35].

The range of diffusion processes that can occur on a network includes the spread of epidemics [9], [10] and information [14], viral marketing [12], [13], word-of-mouth recommendation [11], money exchange, e-mail forwarding [36], and Web surfing [3], among others. Researchers have developed an arsenal of centrality metrics to study the properties of networks, including degree, closeness [37], graph [38] and betweenness [21]; Markov process-based random measures like the Hubbels model [39]; path-based ranking measures like the Katz score [25], SenderRank [26], and eigenvector centrality [22]. However, as Borgatti noted [1], most centrality measures make implicit assumptions about the diffusion process occurring on a network. In order to give correct predictions, these assumptions must match the actual dynamics of the network. Borgatti classified dynamic processes according to the trajectories they follow (geodesic, path, trail, walk) and the method of spread (transfer, serial or parallel duplication). We on the other hand maintain that a simpler classification scheme, that divides dynamic processes into conservative and non-conservative, captures the essential differences between them and informs the choice of the centrality metric. Apart from PageRank and Alpha-Centrality, other measures can be classified as conservative or non-conservative.

Online social networks provide us the unique opportunity to study the dynamic processes occurring on networks. Some studies compared empirical measures, such as tweets and mentions on Twitter [28], [40], with centrality metrics including PageRank and in-degree centrality. We on the other hand, differentiate between the two distinct methods of quantifying influence: *estimating* influence by measuring dynamics of social network behavior and using centrality metrics to *predict* influence. In addition, we evaluate the predictive influence models using the empirical measurements.

Similar to personalized PageRank [4] for conservative diffusion, each user's unique notion of importance in non-conservative diffusion can be captured using customized starting vector for individual users in Alpha-Centrality, leading to personalized Alpha-Centrality. The use of residual vectors and incremental computation in the calculation of approximate Alpha-Centrality leads to scalability of the method. Moreover, as in personalized PageRank, these residual vectors can be shared across multiple personalized views, scaling the personalized Alpha-Centrality metric. Analogous to approximate PageRank [2], in approximate Alpha-Centrality, at each iteration residual vector is redistributed to reduce the difference between the Alpha-Centrality vector and its approximate version. However, the process of redistribution of the residual vector mimics the kind of diffusion the model emulates. For approximate in PageRank, the redistribution of residual vectors is conservative (with the total weight of the residual vector conserved). On the other hand, in approximate Alpha-Centrality, the redistribution of residual vectors is not conservative.

VII. CONCLUSION

We described two fundamentally distinct diffusion processes, which can be mathematically differentiated based on whether or not they conserve the quantity that is diffusing on the network. Random walk, which conserves the probability density of the diffusing quantity, can be modeled as a conservative diffusion process, while epidemics and information spread can be modeled as non-conservative diffusion process. We showed that centrality metrics, such as PageRank and Alpha-Centrality, can be classified as conservative or non-conservative based on the implicit assumptions they make about the redistribution of weight. We showed that since Alpha-Centrality is mathematically equivalent to non-conservative diffusion, it should be used to identify central nodes in online social networks whose primary function is to spread information, a non-conservative process. Future work includes applying this analysis to other online social networks like Twitter and exploring how diffusion process affect other aspects of social network analysis. Our work provides just the initial study of non-conservative diffusion — much work has to be done to understand its properties and extension, for example, application to personalized Alpha-Centrality may be productive. We hope that our work motivates readers to study the properties of non-conservative diffusion and investigate the use of non-conservative in social network analysis.

REFERENCES

- [1] S. Borgatti, "Centrality and network flow," *Social Networks*, vol. 27, no. 1, pp. 55–71, January 2005. [Online]. Available: <http://dx.doi.org/10.1016/j.socnet.2004.11.008>
- [2] R. Andersen, F. Chung, and K. Lang, "Local graph partitioning using pagerank vectors," in *Proc IEEE Foundations of Computer Science*, 2006, pp. 475–486.

- [3] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Stanford Digital Library Technologies Project, Tech. Rep., 1998. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1768>
- [4] G. Jeh and J. Widom, "Scaling personalized web search," in *Proceedings of the 12th international conference on World Wide Web*, ser. WWW '03. New York, NY, USA: ACM, 2003, pp. 271–279. [Online]. Available: <http://dx.doi.org/10.1145/775152.775191>
- [5] H. Tong, C. Faloutsos, and J.-y. Pan, "Fast Random Walk with Restart and Its Applications," in *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, December 2006, pp. 613–622. [Online]. Available: http://www.cs.cmu.edu/~hjtong/pdf/ICDM06_tong.pdf
- [6] S. Fortunato and A. Flammini, "Random Walks on Directed Networks: the Case of PageRank," *International Journal of Bifurcation and Chaos*, vol. 17, pp. 2343–2353, Sep 2007. [Online]. Available: <http://arxiv.org/abs/physics/0604203>
- [7] E. M. Rogers, *Diffusion of Innovations*, 5th Edition. Free Press, August 2003. [Online]. Available: <http://www.worldcat.org/isbn/0743222091>
- [8] L. M. A. Bettencourt, A. Cintrón-Arias, D. I. Kaiser, and C. Castillo-Chávez, "The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models," *Physica A: Statistical Mechanics and its Applications*, vol. In Press, Corrected Proof, Jun 2005. [Online]. Available: <http://dx.doi.org/10.1016/j.physa.2005.08.083>
- [9] R. M. Anderson and R. May, *Infectious diseases of humans: dynamics and control*. Oxford University Press, 1991.
- [10] H. W. Hethcote, "The Mathematics of Infectious Diseases," *SIAM REVIEW*, vol. 42, no. 4, pp. 599–653, 2000.
- [11] J. Goldenberg, B. Libai, and E. Muller, "Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth," *Marketing Letters*, pp. 211–223, August 2001. [Online]. Available: <http://www.complexmarkets.com/files/TalkofNetworks.pdf>
- [12] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," 2003. [Online]. Available: <http://citeseer.ist.psu.edu/kempe03maximizing.html>
- [13] J. L. Iribarren and E. Moro, "Impact of Human Activity Patterns on the Dynamics of Information Diffusion," *Physical Review Letters*, vol. 103, no. 3, pp. 038 702+, Jul 2009. [Online]. Available: <http://dx.doi.org/10.1103/PhysRevLett.103.038702>
- [14] K. Lerman and R. Ghosh, "Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks," in *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010. [Online]. Available: <http://arxiv.org/abs/1003.2664>
- [15] M. Newman, A.-L. Barabasi, and D. J. Watts, Eds., *The Structure and Dynamics of Networks*. Princeton, NJ, USA: Princeton University Press, 2006.
- [16] P. Bonacich, "Power and centrality: a family of measures," *The American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.
- [17] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, "Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint," *Reliable Distributed Systems, IEEE Symposium on*, vol. 0, pp. 25+, 2003. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.129.2340>
- [18] G. V. Steeg, R. Ghosh, and K. Lerman, "What stops social epidemics?" in *submitted to Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Feb 2011.
- [19] N. Bailey, *The Mathematical Theory of Infectious Diseases and its Applications*. London: Griffin, 1975.
- [20] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical Review Letters*, vol. 86, no. 14, pp. 3200–3203, 2001.
- [21] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.
- [22] P. Bonacich and P. Lloyd, "Eigenvector-like measures of centrality for asymmetric relations," *Social Networks*, vol. 23, no. 3, pp. 191–201, 2001.
- [23] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.
- [24] R. Ghosh and K. Lerman, "A Parameterized Centrality Metric for Network Analysis," *Physical Review Letters E*, 2011.
- [25] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, March 1953. [Online]. Available: <http://dx.doi.org/10.1007/BF02289026>
- [26] C. Kiss and M. Bichler, "Identification of influencers-measuring influence in customer networks," *Decision Support Systems*, vol. 46, no. 1, pp. 233–253, 2008.
- [27] K. Lerman and T. Hogg, "Using a Model of Social Dynamics to Predict Popularity of News," in *Proceedings of 19th International World Wide Web Conference*, 2010. [Online]. Available: <http://arxiv.org/abs/1004.5354>
- [28] M. Cha, H. Haddadiy, F. Benevenutoz, and K. P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010.
- [29] C. Lee, H. Kwak, H. Park, and S. Moon, "Finding Influentials from Temporal Order of Information Adoption in Twitter," in *Proceedings of 19th World-Wide Web (WWW) Conference (Poster)*, 2010.

- [30] E. Katz and P. Lazarsfeld, *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Transaction Publishers, October 2005. [Online]. Available: <http://www.worldcat.org/isbn/1412805074>
- [31] J. J. Brown and P. H. Reingen, "Social Ties and Word-of-Mouth Referral Behavior," *The Journal of Consumer Research*, vol. 14, no. 3, pp. 350–362, 1987. [Online]. Available: <http://dx.doi.org/10.2307/2489496>
- [32] D. J. Watts and P. S. Dodds, "Influentials, Networks, and Public Opinion Formation," *Journal of Consumer Research*, vol. 34, no. 4, pp. 441–458, December 2007. [Online]. Available: <http://dx.doi.org/10.1086/518527>
- [33] M. Trusov, A. V. Bodapati, and R. E. Bucklin, "Determining Influential Users in Internet Social Networks," *Journal of Marketing Research*, vol. XLVII, pp. 643–658, 2010.
- [34] P. Boldi, M. Santini, and S. Vigna, "Pagerank as a function of damping factor," in *Proc. of the 14th International World Wide Web Conference*, 2005, pp. 557–566.
- [35] F. Chung and W. Zhao, "Pagerank and random walks on graphs."
- [36] D. Liben-Nowell and J. Kleinberg, "Tracing information flow on a global scale using Internet chain-letter data," *Proceedings of the National Academy of Sciences*, vol. 105, no. 12, pp. 4633–4638, March 2008. [Online]. Available: <http://dx.doi.org/10.1073/pnas.0708471105>
- [37] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, pp. 581–603, 1966.
- [38] P. Hage and F. Harary, "Eccentricity and centrality in networks," *Social Networks*, vol. 17, pp. 57–63, 1995.
- [39] C. Hubbel, "An input-output approach to clique identification," *Sociometry*, vol. 28, pp. 377–399, 1965.
- [40] C. Lee, H. Kwak, H. Park, and S. Moon, "Finding influentials from temporal order of information adoption in twitter," in *19th World-Wide Web (WWW) Conference Poster*, 2010.
- [41] F. Gebali, "Markov chains," *Analysis of Computer and Communication Networks*, p. 65:122, 2008.

APPENDIX

Replication matrix \mathcal{W}_n can be written in terms of its eigenvalues and eigenvectors as:

$$\mathcal{W}_n = X \Lambda X^{-1} = \sum_{i=1}^{|V|} \lambda_i Y_i \quad (16)$$

where X is a matrix whose columns are the eigenvectors of \mathcal{W}_n . Λ is a diagonal matrix, whose diagonal elements are the eigenvalues, $\Lambda_{ii} = \lambda_i$, arranged according to the ordering of the eigenvectors in X . Without loss of generality we assume that $\lambda_1 > \lambda_2 > \dots > \lambda_n$. The matrices Y_i can be determined from the product

$$Y_i = X Z_i X^{-1} \quad (17)$$

where Z_i is the selection matrix having zeros everywhere except for element $(Z_i)_{ii} = 1$ [41]. Therefore

$$\begin{aligned} S(\alpha, t) &= \sum_{k=0}^t (\alpha \mathcal{W}_n)^k \\ &= I + \alpha \lambda_1 \sum_{i=1}^n \frac{(-1)^{\mathcal{I}_i} (1 - \alpha^{t+1} \lambda_i^{t+1})}{(-1)^{\mathcal{I}_i} (1 - \alpha \lambda_i)} Y_i \end{aligned} \quad (18)$$

where $\mathcal{I}_i = 0$ if $\alpha |\lambda_i| < 1$ and $\mathcal{I}_i = 1$ if $\alpha |\lambda_i| > 1$. As obvious from above, for Equation 18 to hold non-trivially, $\alpha \neq \frac{1}{|\lambda_i|} \forall i \in 1, 2, \dots, n$. Now assuming $|\lambda_1|$ is strictly greater than any other eigenvalue

$$S(\alpha, t) \approx I + \frac{(-1)^{\mathcal{I}_1} (\alpha \lambda_1 (1 - \alpha^{t+1} \lambda_1^{t+1}))}{(-1)^{\mathcal{I}_1} (1 - \alpha \lambda_1)} Y_1.$$

For any matrix M , let $\|M\|_1 = \sum_{i,j} M[i, j]$. Therefore, the expected number of paths is $\|S(\alpha, t)\|_1$. The expected path length is given by:

$$\begin{aligned} \frac{\sum_{k=0}^t k \alpha^k \|\mathcal{W}_n^k\|_1}{\sum_{k=0}^t \alpha^k \|\mathcal{W}_n^k\|_1} &= \frac{\alpha \frac{d\|S(\alpha, t)\|_1}{d\alpha}}{\|S(\alpha, t)\|_1} \\ &\approx (-1)^{\mathcal{I}_1} \left(\frac{1}{1 - \alpha \lambda_1} - (t+1) \frac{\alpha^{t+1} \lambda_1^{t+1}}{1 - \alpha^{t+1} \lambda_1^{t+1}} \right) \end{aligned}$$

Therefore, as $t \rightarrow \infty$ and $\alpha |\lambda_1| < 1$, the expected path length is approximately $\frac{1}{1 - \alpha \lambda_1}$, and for $\alpha |\lambda_1| > 1$ it is $O(t)$.